Psychometric Audit and Technical Evaluation of Pearson's and Internet Testing Systems' Operational Examination Processing Systems and Procedures for the Mississippi Subject Area Testing Program

Richard M. Luecht, Ph.D.

Terry A. Ackerman, Ph.D.

Center for Assessment and Research Technology

Greensboro, North Carolina

20 August 2012

This report reflects interpretations of events, opinions and recommendations by the authors as independent auditors and senior technical staff at the Center for Assessment and Research Technology.

1.0 Introduction

Pearson develops and administers four end-of-course examinations as part of the Mississippi Subject Area Test Program, 2nd Edition (SATP2) for the Mississippi Department of Education (MDE): (i) Algebra I; (ii) Biology I; (iii) English II; and (iv) U.S. History. All four examinations must be passed for students to graduate from high school in Mississippi. Internet Testing Systems (ITS) is the subcontractor for the online versions of the examinations. The online version is only administered to students who fail the corresponding paper-and-pencil version.

Due to an apparent breakdown in the QC procedures associated with answer-key validation in a legacy Pearson examination processing system[1] and related QC deficits in the test content and response handling in ITS' online delivery system, an examinee-response mapping error was not detected in successive years going back to 2008. The mapping error caused over 100 students who should have received a passing score on the SATP2 Biology examination to instead receive a failing score. Pearson acknowledged the scoring and reporting error and isolated the apparent cause. In addition, Pearson has now implemented various test content reviews and quality assurance (QA) safeguards—automated and manual—that are intended to prevent similar issues from occurring in the future.

MDE authorized the Center for Assessment and Research Technology (CART) to carry out a psychometric audit and technical evaluation of Pearson's and ITS's information processing systems and operational procedures for test assembly and composition, item and rubric/answer-key validation, and associated quality control processes germane to maintaining score integrity for the MDE SATP. In addition to an extensive and detailed review of documentation surrounding the scoring error and associated item-response mapping problems, CART staff also visited Pearson's San Antonio offices on 24 July 2012 and met with Pearson program staff for the MDE SATP2 and Assessment and Information Quality (AIQ) staff, as well as a senior ITS executive.

This report summarizes CART's psychometric audit and technical review. Section 2.0 describes the key causal details underlying the scoring error event and outlines the QC/QA steps undertaken by Pearson to eliminate problems and ideally prevent other potential examination processing errors in the future. Section 3.0 provides an evaluation of Pearson's proposed solution and prevention measures from the perspective of "best practices" in examination data management design and psychometric data QC analyses. Specific recommendations for improvement are included in Section 3.0.

[1] The legacy system was originally developed by Harcourt. Pearson acquired Harcourt in a corporate merger.

2.0 Details and Root Cause of the Scoring Error

This section has two parts. Section 2.1 details the apparent cause-and-effect of the response data-mapping problem discovered in 2012 for online version of the SATP2 Biology I examination. Section 2.2 lays out Pearson's list of proposed QA/QC corrective action steps.

2.1 Details of Pearson's Root Cause Analysis

Pearson has adopted a formalized set of procedures for documenting examination handling and processing errors called a *root cause analysis* (RCA). RCA refers to any formalized or structured approach for identifying as specifically as possible the conditions and/or factors that result in negative or harmful consequences. In addition to describing the cause(s), RCA ideally also allows organizations to identify the procedures, behaviors, actions, inactions, or conditions that need to be modified to prevent recurrences of similar harmful outcomes. In this sense, although RCA is clearly reactive to a particular problem like the SATP2 scoring error at the onset, it is predicated on the assumption that future problems are best *proactively* solved by attempting to address, correct or eliminate the root causes of past systemic or specific design or operational failures—i.e., to adopt a *preventative* strategy for the future.

A RCA (Pearson tracking number RCA253877) was initiated in March 2012 when Pearson encountered a scoring error in the online SATP2 Biology I retest exam. The error was related to the underlying data-referencing (mapping) of examinee response choices for a small number of multiple-choice (MC) items. Despite the apparent scope of the error being limited to only three MC items used on the online form, the matter remains serious for three reasons. First, the error was ultimately determined to have occurred in multiple years going back to 2008. That should not have happened and potentially represents a QC failure that can seriously erode the necessary trust between a testing services vendor like Pearson and the client—in this case, the MDE. Second, it now appears that the error could have been detected and likely remedied, but was instead assumed to be a simple human keyboarding error during "master testing" of the online test form. As a result, no action was taken to remedy the matter and the error persisted over multiple years. Third, and most importantly, the failure of the operational processing procedures and systems to detect the item-response mapping errors resulted in 126 examinees incorrectly receiving a failing score on the end-of-course SATP Biology I examination.

As part of the computer-based test-form publishing process, items are routinely sent by Pearson to ITS in prescribed digital formats. The items and associated digitized test materials are then reformatted as needed to XML and/or browser-readable HMTL and compiled into an encrypted "resource file" for delivery via ITS' online testing system. The

resource file contains all of the item data, the presentational test-delivery controls, and the test form data needed to administer a test, including all relevant graphics/images and other exhibits. Ideally, the most current version of each resource file is complete locked down to ensure the integrity of the item data and rendered items—that is, preventing any changes without going through a strict set of protocols. While the data lock-down is essential, it has one drawback. If we assume that the locked down item and test-form data are error free, the lock-down mechanism preserves that state. Conversely, if there are errors—as happened in this case across multiple years—they are likewise preserved until corrected.

In this particular case, web-ready graphics provided by Pearson to ITS in late 2007 for the initial import of three items were determined by a test production team at ITS to be inappropriate for rendering on a typical computer screen. A graphics editor on the ITS test production team regenerated the illegible graphics from the [imported] encapsulated postscript (EPS) files as larger, more readable images. The subject items were #3002127 (sequence #20) and #13228 (sequence #41). The editor also recreated the options in HTML format for item #10732 (sequence #36). During this graphics editing and reformatting process the editor determined that the positions of the "B" and "C " MC option values in the table layout appeared to be wrong. These item-specific options-positioning problems were then corrected by relocating the graphics into different "cells" within the virtual option table. However, the underlying item-specific, numerical option-sequencing values (1, 2, 3, and 4) used for mapping each examinee's selected-response choices to the on-screen options were not correspondingly updated. In particular, the improper option and choice sequencing the middle options ("B" and "C") for item #3002127 (form L2) directly impacted the scores for126 examinees who were within one point of the passing cut score. Although the same data-referencing problem occurred for the other two items, there was no direct impact on students' scores or pass/fail decisions.

As alluded to earlier, the data-referencing or mapping error should have been caught through "master testing". Under the usual approach to master testing, small batches of answer sheets[2] containing prescribed response such as all correct answers or particular response patterns such as all of the same responses (e.g., all "A" choices selected) are fed into the system. In the context of using standardized multiple-choice "bubble" style answer sheets, each desired response pattern is directly marked on the answer sheet. The answer sheets are identical to the sheets the actual examinees complete. The various "master" answer sheets are then scanned and run through the system to evaluate potentially miskeyed items or scanning errors that result in response score patterns that do not match expectations based on the scanned inputs.

[2] Master testing was originally developed for scanned answer sheets. The extension of the method for computer-based testing is rather straight-forward (Luecht, 2005b, 2012).

ITS apparently did carry out a crude type of manual mastery testing for the online SATP2 examinations. They had staff members manually enter the desired response patterns using the keyboard to enter the prescribed response choices. The keyboard-entered choices apparently properly mapped to the expected responses and failed to detect any problems. And, if problems were detected, the content person carrying out the review may have assumed that the discrepancy was due to simple keyboard entry error during the manual data entry phase of the master testing process. The actual mapping problem occurred when students used the mouse to select their choices on-screen. Here, the stored response representation was wrong for mouse-selected choices of "B" or "C" on the three items noted above. Furthermore, the item-response mapping problem only affected the students' scores for one of the three items.

A routine item analysis or IA (e.g., Allen & Yen, 2001; Luecht, 2005b) would likely have also detected the problem by generating two different statistical warning flags: (i) a very low or negative item-total correlation signaling that proportionally more of the better performing students were getting the item incorrect than lower performing students[3]; and (ii) a potentially miskeyed item flag generated whenever a non -keyed MC response option has a positive response-total score correlation. This type IA was not performed for the online SATP2 examinations.

The problem item (#30002127) was used operationally from 2008 to 2012. However, the error was first discovered in March 2012 on the online L2 test form. The error appears to have impacted 126 students who would have gotten the item correct under proper response mapping, but who ultimately failed the examination because of the mapping error.

2.2 QC Proactive and Preventative Steps Proposed by Pearson

Obviously, no testing vendor intentionally or negligently treats even minor examination processing errors as "acceptable", despite their scope or impact. That was certainly <u>not</u> the case for Pearson or its online test delivery subcontractor, ITS. There is no causal "motive" to investigate for this type of error—merely a set of causal circumstances that happened and now must be corrected. Ideally, the corrective actions will prevent future errors.

[3] Examinee performance is usually determined for purposes of computing a product-moment correlation by aggregate performance on all other test items. In this context, the apparent performance range restriction due to having only previously failing students sit for the online version of the SATP2 Biology I examination may have reduce the likelihood of a negative correlation, somewhat. However, that potential lack of flagging power for subgroups in a population is not sufficient reason to avoid routine item analyses for answer key validation purposes.

Pearson's first step in handling any examination processing error is to begin a comprehensive RCA (see Section 2.1) and assemble an investigative team. For the SATP2 problem, a the response team was comprised of State Services program staff for the MDE SATP2, psychometricians involved with results processing and analysis for these examinations, staff members from the Assessment and Information Quality (AIQ) area, computer programmers and Information Technology staff, and appropriate technical and administrative staff from Internet Testing Systems (ITS). The formal initiation of a RCA at Pearson also automatically starts a detailed documentation of the problem(s) and a corrective action trail that is entered into a tracking system called the Corrective and Preventative Action (CAPA) workflow. Pearson was very cooperative in sharing with CART staff information about the RCA and details from the CAPA process.

Pearson has proposed an action plan comprised of six QA/QC-related steps that focus on improved production control over the online SATP2 items and related test materials and enhanced documentation and communication within Pearson for all individuals performing all item authoring, editorial and test development, data management, and examination processing for the SATP2 examinations. The six proposed steps are as follows (source: Pearson RCA253877).

1. Establish a formal communication with ITS stating that Pearson will provide web-ready graphics to ITS.
2. Establish a formal communication with Pearson that no creation of replacement graphics will be performed by ITS.
3. Establish a formal communication with Pearson that manual manipulation will be recorded in the ITS log and Pearson must validate all changes.
4. Software will be enhanced to show option values during testing for content review completed by ITS.
5. Pearson standard process will be documented and must strictly be followed.
6. Pearson will internally review the existing process and document chronological order of opportunities for correction.

In addition, Pearson has already rewritten their training manual, established a policy limiting the number of individuals authorized to perform the key validations and review of results, retrained key staff, and updated all relevant documentations.

It should be noted that there are two additional, mitigating factors that could potentially complicate the implementation of these action steps. First, because the MDE SATP2 contract and all examination processing was originally managed by Harcourt—a company subsequently acquired by Pearson—there may be some large-scale system integration issues related to software, database design, and operational processing due to differences between the legacy Harcourt systems and Pearson's complex examination

processing and tracking systems.. This point is not made to gloss over Pearson's legal and professional responsibilities to provide a seamless transition of services and outcomes for its clients; it merely suggests that the full integration of Harcourt's legacy systems with Pearson systems could potentially limit implementation of systemic solutions—even solutions and enhancements already verified through acceptance testing for other examination programs managed by Pearson. Second, ITS is an independent subcontractor with its own item banking, online test delivery and examination data management systems and operational procedures. The subcontractor has apparently corrected the specific problem on the Biology I L2 form used in March 2012 (and for any future test administrations). More specifically, the ITS test production team analyzed the response coding problems and confirmed that there were no other items with this issue in the L2 Biology form, and that the issue with the three specific items has existed in production since the tests were originally published in 2008. ITS further confirmed that there are no other items with similar problems on any other MDE SATP2 test forms. Pearson's AIQ group apparently also verified that the problems with the response coding for the three items described in Section 2.1 were resolved insofar as reviewing the SATP2 staging and production environments. However, these remedies do not suggest any systemic changes to ITS' data management systems or operational procedures, beyond these specific forms.

3.0 Evaluation and Recommendations

CART successfully carried out a thorough psychometric and systems audit of Pearson's and-to-end assessment system operations for another state testing program in 2010. That Pearson audit experience, combined with CART's extensive operational experience with respect to large-scale systems design issues, examination processing, QC, and psychometric analysis, allows CART to objectively and competently evaluate the MDE issue, the RCA conclusions and the adequacy of the corrective actions and associated preventative measures proposed by Pearson.

Our evaluation emphasizes QA/QC in the same spirit as the Pearson RCA procedures. However, our evaluation also emphasizes what we consider to be testing industry best practices in designing automated early warning systems with mandatory follow-up and robust database structures for item, test and examinee data that are somewhat immune to errors such as the mapping problem that occurred here. Pearson already has many of these features in place and is working on better automation throughout their organization. However, legacy systems still in use for particular testing program may pose a serious hindrance that progress.

We also stress the need for competent version controls with "lock downs" and strict policies about who can makes changes, documented policies as to why the changes are made, how the changes are made, the magnitude of changes allowed, who logs the changes,

who verifies the changes, how the changes are tracked, and ultimately, how the most recent, acceptance-tested version of any data components—that is, items, test forms, graphics, tables, exhibits and manipulables, and examinee data record—software, actions or procedures are guaranteed to be thoroughly and robustly disseminated and implemented. In addition, despite "lock downs" of all approved and acceptance-tested data, test materials and software, CART believes that it is a fundamental psychometric best practice to engage in routine monitoring and ongoing software or human-initiated QC and analytics and that ensure and then preserve the integrity of all examination data. It is up to the MDE and Pearson as to whether those recommendations can or should be followed.

For the record, Pearson actually gathers and tracks enormous amounts of QC data/metrics throughout virtually every aspect of their processes. Real-time data are gathered and provided to executives, managers, and employees throughout the Pearson organization, including standardized metrics such as defects-per-million-opportunities, on-time examination starts, cost of quality, and other metrics that focus on performance and load systems testing/scalability. That said, all of these QC systems, data and metrics failed to predict the problem with the MDE SATP2 examinations and further failed to even detect the problem through almost four years of operational use of the Biology I retest examination. At the very least, the facts indicate such that there is still room for improvement with respect to Pearson's ongoing QA/QC efforts.

Our evaluation is summarized in three parts. Section 3.1 evaluates the facts surrounding the MDE SATP2 scoring error incident and associated examinee response-handling issues. Section 3.2 evaluates the adequacy and concreteness of the corrective actions and preventative steps proposed by Pearson. Finally, Section 3.3 lays out CART's recommendations for improvement and considerations of related issues that might need to be addressed for the SATP2 with respect to both the online and paper-and-pencil tests.

3.1 Evaluation of the SATP Incident

Pearson's initiation of the RCA process seems reasonable, once the scoring error was actually detected. Obviously, the fact that it took four years for Pearson and ITS to detect the item-response mapping error and scoring problem was highly unfortunate, especially considering that error detection may have been relatively straightforward using standard item analysis mechanisms and statistical results[4].

The authors reviewed all of the documentation provided by Pearson[5] and engaged in a series of email correspondences and conference calls with Chris Skapyak, Vice

[4] It is still not absolutely clear when key validations were performed for Form L2 or how any available results may have been used to evaluate the quality of the items and signal any answer-key or other problems.
[5] All data were made available to CART via a secure FTP server.

President of Quality and Continuous Improvement at Pearson. As part of the audit and technical evaluation, a senior member of the CART technical staff, Dr. Terry Ackerman, also visited the Pearson facilities in San Antonio on 24 July 2012. Dr. Ackerman met Melinda Orta from the Pearson MDE Program Team, Brandon Burgess, Rebecca McCully and Joe Magargee from the Pearson AIQ group, and Lisa Ward from ITS. The meeting was extremely useful. It focused on an overview of the SATP2 program and Pearson's work with MDE, Pearson's AIQ procedures in general for the Mississippi testing program and specific to the SATP2 problem, ITS' explanation of the SATP2 problem and a review of their examination QA procedures.

The fundamental problem appears to have centered on there being far too much latitude extended to the ITS graphic editor and test production staff, in general. From a pure data management perspective, "version 1.0" of an item begins when the first instance of the item is authored. The image of the item, any interactive functionality, associated graphics, tables, reference materials, auxiliary tools (e.g., online calculators), and data including content codes, answer keys, rubrics, and psychometric statistics should considered to be "locked down" from a data integrity perspective. Pearson has many software and database controls to track and ensure the integrity of the items within their item data base. It is not clear what then happens—which types of modifications are allowed or what types of data policies and controls are in place as to who can make the changes, how they are logged and acceptance-tested, etc., once an item leaves the Pearson environment and is passed onto a test delivery like ITS.

In this case, an ITS graphics editor made what appears to be an aesthetic decision to modify (recreate in a larger size) some specific graphics used as static exhibits for a small number of items. The unlogged changes for three items also repositioned some of the response options. These modifications actually introduced *new versions* of the items (i.e., new instances of the data) and should have been: (a) logged as such; (b) subjected to formal acceptance testing and sign-off procedures; (c) locked down as a new version of the item data, using an effective version control mechanism; and (d) updated in Pearson's master item database, ITS' item databases and all current resource file versions. The fact that the incident was an exhibit (i.e., graphic) rather than item text or the usual types of CBT radio controls text or tagged XML content used with most online tests for MC items is incidental to the broader need for documentation and controlled dissemination of any and all changes that affect items, even in seemingly trivial ways.

From a certain perspective, Form L2 of the online Biology I exam actually was "locked down" insofar as reusing that the same form of the test, without modification, from 2008 to 2012. However, that data management rationale only works if an error-free version of the test were lock-down rather than perpetuating an existing error (also, see Footnote #4).

Should the error have been detectable? The simple answer is "probably"—via either master testing (MT) and/or item analysis (IA). MT, as briefly described earlier in this report, was originally developed for carrying out quality control of scanned answer sheets. Master answer sheets are bubbled in according to defined protocols of interest. For example, if all correct answers are keyed, the response string should return a vector of all ones indicating perfect performance on the test form. Other patterns of responses are often also fed through the scanning system (e.g., all "A" options selected, etc.).

A somewhat rudimentary form of MT was apparently carried out by ITS by having a staff member manually type in the desired patterns of responses using the computer keyboard[6]. This type of manual QA procedure was problematic in two ways. First, a different response input mechanism was used for manually master testing the L2 form than the examinees typically used during their online examination session. During the online examination, students use the mouse to click on a response option, hot-spot, or other selectable choice. The ITS test delivery engine then maps each selection to an encoded response option (e.g., "2" for the second on-screen option). However, only the final, encoded response selection is stored for each student-by-item response *transaction*, if the student decides to change an earlier response. In contrast, during the ITS MT process, the responses were manually entered using the computer keyboard. This allowed the master tester to directly record the scripted option selections. However, it also completely circumvented the response handling glitch.

The second method of detection that should have been in place for the SATP2 would have involved routinely running standard psychometric IA procedures for every test administration and/or period of interest. Most IAs are relatively quick and inexpensive to run using customized or commercial software and can easily be run even with small samples of examinee data and regardless of the nature of the examination (i.e., whether for a regular administration, retest, or for special accommodations test forms such as extended time or large-print editions). Although IAs may have been run for the regular, paper-and-pencil versions, it is not clear that any follow-up IAs were run from 2008 to 2012 for the online L2 form. Furthermore, if the data from the paper-and-pencil and online version were combined, the larger sample taking the "error-free" forms of the regular paper-and-pencil test would have swamped the IA results and probably covered up any problems.

There are obvious cost and contractual issues associated with who runs the IAs and how they are paid for those activities. Such issues exceed the scope of this audit and technical evaluation report. Nonetheless, it is arguably best practice for a testing company

[6] A more efficient automated variant of this procedure would be to prepare response decks and load them into to online system using the exam restart option—that is, the capability to restart an exam instance that crashes. However, that more automated approach would still not have led to the detection of the SATP2 error.

to: (a) always offer the service and (b) make sure that the client understands the costs and potential QC benefits.

Would a psychometric IA have caught the problem item? Again, the answer is, "perhaps." For example, item point-biserial correlations are routinely computed as part of any psychometric IA. A point-biserial is a product-moment correlation between a dichotomous variable (e.g., a MC item response scored 0=incorrect or 1=correct) and a continuous variable such as a total test score. Assuming that a larger proportion of the higher performing students, based on their total test scores, should correctly answer well-designed items than lower performing students, a near-zero or negative point-biserial correlation would indicate a reversal from expectation and flag the item for review[7]. Point biserial correlations can also be used to correlate *incorrect* response selection (i.e., multi-nominal response selections) with total scores. This type of analysis is typically part of a MC "distractor analysis" (Luecht, 2005b). A positive correlation between an incorrect response and the total score could signal a potential second correct answer or the most plausible correct key for a miskeyed item. In any case, these types of statistical flags would signal a potential problem. Mandatory item reviews, follow-up actions and signoffs would still be required to ensure than all flagged items were responsibly handled and any key or response handling issues fully resolved before scoring.

3.2 Evaluation of Pearson's Corrective Actions and Preventative Steps

This section evaluates the appropriateness and potential effectiveness of Pearson's six proposed corrective action steps. The proposed corrective steps focus on *prevention* through better documentation and communication, retraining of staff, and isolation of particular editorial procedures (e.g., making item text or graphics changes at Pearson, not ITS). The following paragraphs provide a review and commentary on each of the proposed corrective action steps.

Pearson will provide web-ready graphics. This modification of the item authoring and editing process effectively by-passes ITS altogether and, at the very least, allows Pearson to maintain direct control over the rendered form of all graphics used on the SATP. More detail is required, however, with respect to (a) having independent verification and sign-off on all modifications that are within the production "pipeline"; (b) establishing strict version control and reference, including lock down in the master item database and proper dissemination to the resource file; and (c) logging, verification and sign-off of ANY and ALL

[7] All statistical correlations are susceptible to a phenomena known as "variance restriction". When there is very little in a variable (score or response), the variable will tend to demonstrate a diminished correlation with any other variable. The restricted proficiency of a retest sample of examinees potentially restricts the variance of both total test score and item response scores.

subsequent reformatting of the online items performed by ITS (e.g., modifications of XML data content, presentational XSL style sheets, graphics formats, or HTML editing).

No creation of replacement graphics will be performed by ITS. This corrective action is a corollary to the previous one. The same commentary provided above holds here.

Manual manipulation should be recorded in the ITS log and Pearson must validate all changes. This corrective action is in direct response to the failure by ITS editors to log the graphics modifications that led to the item response mapping error. Frankly, it is not clear who at ITS can (or should) even make changes. What types of changes are (if ever) allowed? How are logs reviewed, checked and finalized for each resource file? Acceptance testing and sign-off at the test form level may be insufficient. Pearson and ITS should document the detailed item and content review procedures and QC metrics in that respect.

Software enhancement to show option values during testing for content review completed by ITS. This seems to be an important software change that will facilitate master testing. Some CART recommendations for how the data and properties of selected-response options are structured and stored are covered in Section 3.3.

Pearson standard process must strictly be followed. This is a policy statement, not a corrective action. How can this standard be enforced? What QA metrics exist to document staff training, qualification-to-engage, retraining and routine monitoring?

Pearson will internally review the existing processes and document chronological order of opportunities for correction. This is a global statement that implies that a set of actions will be undertaken at Pearson. It would appear that a report detailing the "opportunities for correction" will be prepared at some point in the future. For example, the ongoing key validation process—including timing and implementation of the statistical analysis, interpretive review of results, and mandatory follow-up steps, data modifications and sign-offs on statistically flagged items at both Pearson and ITS—the latter for items or test form that might still be active—needs to be documented. Some of the IA work could be effectively automated to run on a regular basis. As is, this action is much too vague.

Overall, Pearson's response and direct acknowledgement of the errors and detection failures over multiple years appears to be serious and commendable. It was a subtle and not-easy-to-detect problem. The preventative strategy laid out likewise seems reasonable and desirable, although many of the specific details of the corrective action steps are conspicuously missing. The rather obvious focus appears to be on generating better documentation, retraining and overall communication. Still, that communicative strategy provides the almost unavoidable possibility for waning human attention and motivation that could result in occasional deterioration of quality and associated follow-up activity, especially during peak SATP2 production and/or processing cycles for the SATP2. What

happens when processing deadlines are extremely short or when editors and operators become overwhelmed with additional work during peak examination production or processing times? What happens when routine tasks become monotonous and boring, just because they have been performed over and over without finding errors? For example, acceptance testing results evaluations and sign-offs can become almost *pro forma* in any human-led QC system.

Our recommendations in the next section do not offer any simple answers to the "human factors" elements. The dynamics are simply too complex. Rather, we focus on establishing improved and possibly more robust data structures, automated early warning and system-generated QC procedures with mandatory action steps, and use of templates and/or software-enforced restrictions on critical aspects of the item authoring, test-form assembly and production, delivery, examination results processing. While these recommendations are certainly not guaranteed to eliminate errors, they do reflect what CART and many testing organizations consider to be best practices and should be carefully considered by Pearson, ITS and the Mississippi Department of Education.

3.3 CART Recommendations

CART recognizes that Pearson and ITS have extensive operational experience with large-scale assessment programs like the SATP2. It would be presumptive and probably naïve to assume that we likewise have all the answers. That said, there are certain best practices that arguably can and should be adopted for this and other testing programs. Our recommendations focus on four classes of improvements: (3.3.1) enhanced database and data-object structures for selected response items types; (3.3.2) routine and on-going statistical analyses of active examination titles; (3.3.3) enhanced version control, verification and sign-off assessment data objects as well as procedural software; and (3.3.4) periodic, independent QC audits. Some of these issues obviously have important cost and contractual ramifications.

3.3.1 Enhanced Database and Data Structure Considerations

Although there are many different philosophies, benefits and costs associated with various database designs and data structures/frameworks (relational scheme, object-oriented schemes, etc.), there is a data-integrity principle to which all designs must pay homage. If data ceases to have integrity—that is, stable data types and representations across hardware platforms and operating systems, single source representation in a master repository, and consistency in different contexts and within allowable transformations, functions or coercions such as reporting—it becomes almost useless. Structure is a key aspect for maintaining integrity—more specifically, object-oriented design (OOD) structures. OOD structures are usually to be encapsulated and robust with respect to

changes, transportable across platforms, flexible insofar as fitting into new data use contexts, and scalable/extensible in scope. It may be useful for Pearson and ITS to reconsider the design of all item type templates and the associated item data from a much stronger OOD perspective.

As an example, consider the response-handling problem that occurred with the SATP2. The RCA indicated that the cause the problem involved reformatting by an ITS graphics editor that led to an inconsistency in how mouse-click MC selections were stored. Furthermore, master testing failed to detect the anomalies. In reality, however, the differences in response capturing mechanisms for the online examination and master testing (mouse selection vs. keyboard) masked what may be a more **fundamental data design flaw** in the way that rendered response options are mapped to the encoded values stored as the final "response". The referencing ITS uses appears to be *implicit* (i.e., seriated or position-based for certain item types and screen layouts) rather than explicit. For example, *explicit* referencing would store the visible text, data-object reference, alphabetic letter or number displayed and the value actually stored when a particular option is selected, all as part of the data-object properties for each MC option. A robust set of hierarchically oriented structural data templates would allow different item types to reference different properties as needed. The differences between implicit and explicit referencing are often quite subtle from a data systems design perspective, but absolutely crucial from the standpoint of maintaining data integrity. Implicit referencing is often more efficient, but has less integrity than explicit referencing. In this case, implicit referencing appears to have created a "loop hole" in the system whereby two different response input mechanisms led to two different results, even though the student-examinee and master tester were looking at and responding to exactly the same item. Explicit referencing of the options would have made it extremely unlikely for the mapping error to even occur since the visible response options and underlying values stored would have been properties of the options, themselves, and followed any repositioning.

3.3.2 Versioning and Version Controls for All Data

All assessment data objects should have a "version" property to ensure that the most recent version is always used. The most recent instances of an active item or other data component should always be available for all active test forms and one and only one version should ever be active or capable of being referenced. This principle holds for items, graphical images, tables, exhibits, test modules or sections, test forms, and even entire resource files. It is not sufficient to merely re-identify each item or exhibit version. Strict version coding, sign-offs and lock downs of the acceptance-tested versions ought to be possible.

Pearson does have the capability to lock down files in their primary item database system under their AMS system, including automated version control and enforced of naming conventions for items and meta data, and strong data storage policies that ideally preclude storing or imaging files on local drives (e.g., potentially creating multiple active versions of the same items). Furthermore, standard Pearson naming conventions are supposed to include specific versioning information that can be tracked in the AMS system. However, it is not clear whether these types of version controls exist in any of Pearson's legacy systems, or whether the integrity of the version controls and data lock downs is maintained once an item, graphic or other assessment data object leaves the Pearson environment and transitions to an outside test delivery system like the ITS online system.

Although versioning was not specifically an issue for the Mississippi SATP2, there are actually multiple versions of the same the items and exhibits: one version for the paper-and-pencil tests, another version generated by Pearson for upload to ITS, and a third version or representation within the ITS item bank and resource files. Therefore, the potential exists for versioning to be a problem for programs like the SATP2.

3.3.3 Implementing Additional Statistical QC Procedures

The psychometric literature is replete with numerous statistical IA procedures, item response theory (IRT) data-model fit indices and related residual fit analyses, and other aberrant- or unexpected response indicators. Some of these analyses are available through commercial or shareware packages, others are continually being developed, modified and shared globally in modern programming code libraries like the *R* programming language CRAN mirrors (R Development Core Team, 2012).

What is needed to take advantage of these tools are: (a) standardized data queries for extracting the data on demand and (b) a QC plan to routinely run the queries and psychometric analyses, and to follow-up on any anomalies, however, trivial appearing on the surface. The notion of standardized extracts implies that the queries can be requested whenever needed by psychometricians and automatically formatted for the needed analysis, without intervention by database or other information technology staff. It is also essential to introduce mandatory follow-up or defined anomalies, statistical flags, and other detection triggers.

As discussed earlier, it is not certain that a psychometric IA would have actually detected the response-handling problem and scoring errors. Neither is it clear whether ITS or Pearson actually ran separate pre- and post-administration key validation checks for the existing SATP2 test forms (paper-and-pencil and online versions), and whether this process was repeated for each new test administration. However, if the standardized queries and extracts can be implemented, there is certainly no reason not to routinely take

advantage of these statistical QC mechanisms as a way of continually monitoring the quality of the items and examinee data.

3.3.4 Periodic Independent, QC Audits of the SATP

Our last recommendation is more directed at the Mississippi Department of Education than at Pearson or ITS *per se*. Many state testing programs have discovered that having qualified individuals or external organizations conduct routine, independent audits of many of the procedures carried out by testing services vendors reduces errors and anomalous results and increases confidence by the state regarding their results. These independent audits can range from scaled-down, proportionally sampled audits to full-scale replications of analyses and results. As noted earlier, there are obviously cost and benefits to consider, as well as contractual ramifications with the primary vendors and any external consultants or QC groups (non-disclosure agreements, sharing proprietary software, metadata structures,. providing time data access, etc.).

4.0 References

Allen, Mary J. & Yen, Wendy M. (2001). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole Publishing.

Luecht, R. M. (2005a). Operational Issues in Computer-Based Testing. In D. Bartrum and R. Hambleton (Eds). Computer-Based Testing and the Internet, (pp. 91-114). New York, Wiley & Sons Publishing.

Luecht, R. M. (2005b). Item Analysis. In B. Everitt & D. Howell (Eds), Encyclopedia of Statistics in Behavioral Science. West Sussex, UK: John Wiley & Sons, Ltd.

Luecht, R. M. (2012). Operational CBT Implementation Issues: Making It Happen. In R. Lissitz & H. Jiao (Eds.), *Computers and Their Impact on State Assessments: Recent History and Predictions for the Future*. Baltimore, MD: Information Age Publishers.

R Development Core Team (2012). The R Programming Project. [Computer Program]: Authors (http://www.r-project.org/)